Mutational analysis identifies therapeutic biomarkers in inflammatory bowel disease-associated colorectal cancers

Title: Mutational analysis identifies therapeutic biomarkers in inflammatory bowel disease-associated colorectal cancers

Authors:

Shahida Din [α], Consultant Gastroenterologist, NHS Lothian, Gastrointestinal Unit, Western General Hospital, Edinburgh, Scotland, UK.

Kim Wong [α], Senior Computational Biologist, Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

Mike F Mueller, Post-doctoral Fellow**,** Division of Pathology, Centre for Comparative Pathology, Edinburgh Cancer Research Centre, Institute of Genetics & Molecular Medicine, Western General Hospital, University of Edinburgh, Edinburgh, Scotland, UK

Anca Oniscu, Consultant Pathologist, NHS Lothian, Department of Molecular Pathology, Laboratory Medicine, Royal Infirmary of Edinburgh, Edinburgh, Scotland, UK

James Hewinson, Principal Research Associate, Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

Catherine J Black, Consultant Pathologist, NHS Lothian, Department of Pathology, Western General Hospital, Edinburgh, Scotland, UK.

Martin L Miller, Group Leader, Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge UK

Alejandro Jiménez-Sánchez, Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge UK

Roy Rabbie, Clinical Research Fellow, Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

Mamunar Rashid, Senior Computational Biologist, Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

Jack Satsangi, Professor of Gastroenterology, Centre for Genomic and Experimental Medicine, University of Edinburgh, Edinburgh, Scotland, UK

David J Adams, Senior Group Leader, Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

Mark J Arends, Professor of Pathology, Division of Pathology, Centre for Comparative Pathology, Cancer Research UK Edinburgh Centre, Institute of Genetics & Molecular Medicine, Western General Hospital, University of Edinburgh, Edinburgh, Scotland, UK

[α] The first two authors contributed equally

Running Title predictive biomarkers in IBD-associated colorectal cancers

KEY WORDS: Crohn's disease; Ulcerative colitis; hypermutation; DNA repair; immunotherapy

Additional Information

Financial Support

This work was supported by the European Research Council Combat Cancer Program, Cancer Research UK and the Wellcome Trust.

SD acknowledges the financial support of NHS Research Scotland (NRS), through NHS Lothian.

AJS was supported by a doctoral fellowship from the Cancer Research UK Cambridge Institute and the Mexican National Council of Science and Technology (CONACyT).

Corresponding author:

Dr Shahida Din, Consultant Gastroenterologist, Gastrointestinal Unit, Level 1 Anne Ferguson Building, Western General Hospital, Crewe Road South, Edinburgh, Scotland, UK, EH4 2XU.

Email: sdin@ed.ac.uk

Contact telephone: 00 44 (0) 7912092918

Conflict of Interest disclosures:

The authors declare no potential conflicts of interest.

Total Word Count 4917

Main Tables 2

Main Figures 5

Abbreviations

CD	Crohn's disease

dMMR	defective mismatch repair
FFPE	formalin-fixed paraffin-embedded
IBD-CRC	inflammatory bowel disease associated colorectal cancer
InDels	insertions and/or deletions
IQR	inter-quartile range
Mb	megabase
MMR	mismatch repair
MSI	micro-satellite instability
MSS	micro-satellite stable
SNVs	single nucleotide variants
TCGA	The Cancer Genome Atlas
TILs	tumor infiltrating lymphocytes
UC	ulcerative colitis

Statement of Translational Relevance

There is an urgent need to identify predictive biomarkers in inflammatory bowel disease-associated colorectal cancers (IBD-CRCs) to individualise the current prevention, surveillance and treatment programmes. This analysis demonstrates that proximal IBD-CRCs have high mutational rates associated with defects in MMR (MLH1 loss) and DNA *POLE* proofreading function. Hypermutation is associated with a predicted higher neo-epitope load, which could be exploited using immunotherapies in selected patients with hypermutated IBD-CRCs. Prospective studies are required to determine whether analysis for loss of MLH1 in surveillance colonic biopsies could discriminate those at increased risk of developing CRC, permitting a more targeted approach for cancer treatment and surveillance. The identification of driver genes in hypermutated IBD-CRCs could also be used to develop therapeutic agents targeting the corresponding molecular pathways. Our comprehensive analysis of the mutational landscape of IBD-CRCs has revealed several novel approaches that may complement and personalise surveillance and treatment programmes.

Abstract:

Purpose: Inflammatory bowel disease-associated colorectal cancers (IBD-CRCs) are associated with a higher mortality than sporadic colorectal cancers. The poorly defined molecular pathogenesis of IBD-CRCs limits development of effective prevention, detection and treatment strategies. We aimed to identify biomarkers using whole exome sequencing of IBD-CRCs to guide individualised management.

Experimental Design: Whole-exome sequencing was performed on 34 formalin-fixed paraffin-embedded primary IBD-CRCs and 31 matched normal lymph nodes. Computational methods were used to identify somatic point mutations, small insertions and deletions, mutational signatures, and somatic copy number alterations. Mismatch repair status was examined.

Results: Hypermutation was observed in 27% of IBD-CRCs. All hypermutated cancers were from the proximal colon; all but 1 of the cancers with hypermutation had defective mismatch repair or somatic mutations in the proofreading domain of DNA *POLE*. Hypermutated IBD-CRCs had increased numbers of predicted neo-epitopes, which could be exploited using immunotherapy. We identified 6 distinct mutation signatures in IBD-CRCs, 3 of which corresponded with known mechanisms of mutagenesis. Driver genes were also identified.

Conclusions: IBD-CRCs should be evaluated for hypermutation and defective mismatch repair to identify patients with a higher neo-epitope load who may benefit from immunotherapies. Prospective trials are required to determine whether immunohistochemistry to detect loss of MLH1 expression in dysplastic colonic tissue could identify patients at increased risk of developing IBD-CRC. We identified mutations in genes in IBD-CRCs with hypermutation that might be targeted therapeutically. These approaches would complement and individualise surveillance and treatment programmes.

INTRODUCTION

Inflammatory bowel disease associated colorectal cancer (IBD-CRC) is an aggressive complication of chronic inflammation accounting for 10-15% of deaths from IBD. IBD-CRC patients are younger and have a higher mortality than those with sporadic colorectal cancer (1). IBD-CRC arises diffusely in chronically inflamed epithelium via low-grade dysplasia evolving to high-grade dysplasia and eventually adenocarcinoma, although it may also arise without these preceding changes. Technological advances in endoscopy have improved the ability to discriminate between normal and pre-cancerous mucosa (1), however, a third of cancers arise in the interval between scheduled colonoscopies, suggesting poor efficacy of surveillance programmes (2). Colonoscopy-directed mucosal samples are analysed for the presence of dysplasia, which remains the gold standard for predicting future cancer risk in IBD when reported by an expert gastrointestinal pathologist.

The poorly defined molecular and genetic pathogenesis of IBD-CRC impairs our ability to detect and treat colorectal cancer effectively. There is an urgent need to identify biomarkers to improve early detection of colorectal cancer and guide individualised therapy (2).

We elected to study primary IBD-CRC since molecular changes identified in the resected cancer specimen are most likely to represent biomarkers of cancer development. We characterised the hypermutator status of the cancers, defined the mutational signatures and underlying biological processes, and explored the potential clinical translation of this work with respect to immune therapy, and improving current surveillance programmes.

MATERIALS and METHODS

Ethical Approval and Case Identification

The study was approved by the Lothian NHS Research Scotland Human Annotated BioResource, which is an NHS Health Research Authority Ethics Committee-approved Research Tissue Bank (www.hra.nhs.uk; 15/ES/0094) for the use off human tissue surplus to diagnostic requirements. The study was conducted in accordance with the Declaration of Helsinki and the guidance from the Human Tissue Authority on the use of human tissue from diagnostic archives. The Lothian (Edinburgh, UK) pathology database was searched for IBD-CRCs using the terms inflammatory bowel disease, Crohn's disease, ulcerative colitis, colon cancer, colorectal and rectal cancer between 1990 and 2011 by the Tissue Bank appointed pathologist and provided the anonymised tissue samples with linked clinical data. Two independent expert GI pathologists (MJA and CJB) and a specialist IBD Consultant Physician (SD) verified the IBD-CRCs following careful review of the patient's anonymised medical records, the histopathological evidence for IBD in previous colonic samples and in the IBD-CRC resection specimen.

Cases were only included from patients with a previous histopathological diagnosis of IBD and where evidence of IBD in the resected colonic specimen was confirmed by both GI specialist pathologists (CJB and MJA) independently. Sporadic adenomas and colorectal cancers which arose in an area of the colon without prior evidence of IBD in that colonic area were excluded. For each case we have clear histopathological evidence of pre-existing IBD at each tumour site and have confirmed that we have selected IBD-associated colorectal cancers.

Formalin fixed paraffin embedded (FFPE) cancer and uninvolved normal lymph node blocks, removed at the time of surgery, were sectioned, H&E stained and used for

further analysis. In total 31 cases (15 Crohn's Disease (CD): 16 Ulcerative Colitis (UC)) with 34 cancers were used in this study.

Clinical Phenotype Data: Statistical Analysis

The 'survival' package within the R software environment (3) was used to generate the Kaplan-Meier survival curves and perform statistical analyses. Categorical clinical phenotype data was analyzed using the Two-tailed Fisher's exact test; a *P*-value of <0.05 was considered significant. No survival data was available for case 2L (hypermutator) which was excluded from the survival analysis. Cases 15G1 and 15G2 were considered as one case for the purpose of statistical comparison, as they appear to originate from the same precursor clone.

Sequencing Data Generation, Processing and Analyses

Nucleic acids were extracted using Qiagen Allprep FFPE DNA extraction kits, and DNA/RNA was quantified using Agilent GeneChips. For WES, DNA was captured using the Agilent SureSelectXT Human All Exon V5 platform following the manufacturer's protocol. Sequencing was performed using the Illumina HiSeq 2000 platform at the Wellcome Trust Sanger Institute. Raw paired-end sequencing reads (75bp) were aligned and a modified version of the reference genome which includes the GRCh37 primary assembly and additional human contigs and viral sequences that reduce the number of reads erroneously mapped to the primary assembly. Further details on data processing, data quality assessment and sequence alignment are available in the Supplementary Materials and Methods. Somatic point mutations and small insertions and/or deletions (InDels) were identified by comparing cancer and matched normal samples using MuTect (v1.1.7) (4) and Strelka (v1.0.14) (5), respectively. BCFtools (v1.3) (6) was used to identify variants in each patient's

germline (normal lymph nodes) relative to the human reference genome described above. The Ensembl Variant Effect Predictor (7) software was used to predict the effect of each variant on protein sequences, and the clinical significance of the changes was predicted by comparison to variants in the ClinVar database (release date 2016-05-31). For genes with multiple transcripts, we report the effect of the variant on the protein derived from the canonical transcript as annotated in Ensembl release 81 for GRCh37 (Supplementary Table S7; Supplementary Materials and Methods). We also provide the predicted effect on all transcripts in Supplementary Table S8. The Sequenza software package (version 2.1.2) (8) was used to identify somatic copy number alterations. For the non-hypermutator cases, MutSigCV (version 1.4) (9), which identifies significantly mutated genes, and dNdScv (10), which identifies genes under positive selection in cancer, were used to identify driver genes. For the hypermutator cases, microsatellite InDels were excluded from analysis with MutSigCV. The microsatellite InDels were analysed with a modified version of MSMutSig (11) obtained from the author, along with required input files. In dNdScv, which is able to analyze both the SNV and InDel mutations together in both hypermutator and non-hypermutator cohorts, two InDel models are available; one model considers the total number of InDel mutations per gene, and the other model considers unique InDel sites per gene (the “unique-sites model”). For the hypermutator cohort, dNdScv was run using both models, and the results were compared. Only the “unique-sites” model was used for the non-hypermutator cohort. Restricted hypothesis testing was performed on the results from dNdScv, using known cancer genes from the Cancer Gene Census version 81 (12). Further details are available in the Supplementary Materials and Methods. Mutation signatures were identified with the Bioconductor package SomaticSignatures (version 2.6.0) (13) using the non-negative matrix factorization (NMF) algorithm. Details of software

parameters, databases used, variant annotation, variant filtering, validation and the identification of driver genes are available in the Supplementary Materials and Methods.

HLA Typing and Neo-epitope Predictions

Human leukocyte antigen (HLA)-I 4-digit typing was performed using the OptiType 1.0 algorithm and neo-epitopes from missense mutations were predicted by mapping the corresponding protein sequences to the human proteome database (version GRCh37.74). Neo-epitopes with a relative percentile rank ≤ 1% for each HLA-I allele were considered binders (additional details are provided in Supplementary Materials and Methods).

Mismatch Repair Immunohistochemistry

Immunohistochemistry for MLH1, MSH2, MSH6 and PMS2 was undertaken on a cancer tissue microarray and staining was independently scored by two pathologists (MJA and AO). Discordant scores were resolved by staining whole tissue sections with additional MLH1, MSH2, MSH6 and/or PMS2 antibodies.

***MLH1* Promoter Methylation Analysis**

MLH1 promoter methylation was analysed using the EZ DNA Methylation Kit Gold (Zymo Research). Cases from patient 32N failed methylation analyses.

Further experimental details are described in the Supplementary Materials and Methods.

RESULTS

To identify biomarkers associated with IBD-CRC, WES was performed on 34 IBD-CRCs and matched normal lymph node pairs. Twenty-nine patients had one cancer, one patient had three primary cancers separated anatomically (32N) and one patient had two cancers in close proximity to each other (15G) (Supplementary Table S1). WES yielded between 45- and 90-fold coverage of the cancer samples and 26- and 93-fold coverage of the normal lymph node samples (Supplementary Fig. 1). When considering protein coding and splice site mutations only, the cancers from 32N had only 6 mutations common to at least 2 of the 3 cancers, while the two cancers from 15G had over 2345 common mutations (22% of the total mutant positions in these cancers).

Somatic Mutation Rates in IBD-CRC

Somatic point mutations and InDels were identified by comparing exome sequences from cancer tissue with those from uninvolved lymph nodes removed at the time of surgery (Supplementary Tables S7 - S8).The 34 cancers were divided into two groups based on distinct somatic mutation rates. There were 24 non-hypermutator cancers with 2.0-7.0 mutations/Mb, and 10 hypermutator cancers with 32.6-171.3 mutations/Mb (Fig. 1A and Table 1). Two of the 10 cases, 15G1 and 15G2, have a somatic mutation burden greater than 100 mutations/Mb and could thus be defined as ultra-hypermutators.

The proportion of hypermutators in our IBD-CRC data set was 27%, or 9/33 (cases 15G1 and 15G2 were counted once, since they appear to originate from the same precursor clone) and was not significantly higher than the 16% present in a cohort of 382 sporadic CRCs downloaded from The Cancer Genome Atlas (TCGA) Data Portal (see Supplementary Materials and Methods) (right sided binomial test, P=.07), nor

the 28% observed in a larger cohort of 619 sporadic CRCs from Giannakis *et al.* (binomial test, *P*=1) (14). The hypermutator cases had elevated levels of both single nucleotide variants (SNVs) and InDels, with the exception of cases 15G1, 15G2 and 6J, which have similar InDel mutation rates to the non-hypermutator cases (Fig. 1A). The median age at diagnosis was not significantly different between the hypermutators (median 66.2 years; IQR 57.4-78.1 years) and non-hypermutators (median 65.8 years; IQR 51.0-77.0 years).

Overall, there was a significant survival difference between the patients with hypermutator cancers and those with non-hypermutator cancers (log rank test, *P*=.04), with the estimated 10-year survival being 75% in the former group as compared with 36% in the latter (Fig. 1B). However, it is likely that additional factors other than mutator status including stage and age influenced patient survival although we were unable to measure the affects of these variables due to the small sample size.

Strikingly, the hypermutator cancers were all located in the proximal colon and none in the distal colon (two-tailed Fisher's exact test, *P*=.004) (Fig. 2).

Mismatch Repair Abnormalities

To characterise the difference in mutational rates, the cancers were analysed for genetic aberrations associated with dMMR. Seven out of the ten hypermutator cancers had a high frequency of InDels (Fig. 1A), which is indicative of MSI, and showed loss of expression of MLH1 and its heterodimeric binding partner PMS2 (Fig. 2; Supplementary Fig. 2). Loss of MLH1 protein expression results in dMMR, leading to increased somatic substitution and susceptibility to cancer (15). Loss of MLH1 expression can be explained by *MLH1* promoter hypermethylation in 5 of these 7 cancers (Fig. 2; Supplementary Fig. S3). One of the two remaining cancers (33W),

had a somatic nonsense mutation within the ATPase domain (R100*) of the *MLH1* gene, resulting in a truncated protein with a predicted loss of function. We did not find other coding mutations that would cause *MLH1* to be biallelically inactivated in case 33W. We did not detect any point or InDel mutations in *MLH1* in the germline or tumour in case 28E, nor epigenetic silencing by promoter hypermethylation that could explain the loss of *MHL1* expression in this case. The normal expression of MLH1 and low level of somatic InDels in the remaining 3 hypermutator cancers, 15G1, 15G2, and 6J, point to alternative mechanisms leading to hypermutation. In contrast, loss of expression, promoter hypermethylation and non-silent somatic mutations were not observed in *MLH1* in the non-hypermutator cases (Fig. 2; Supplementary Table S2B).

Mutations in DNA Polymerase Proof-reading Domains

During DNA replication, DNA fidelity is maintained by the proof-reading function of DNA polymerases. Germline mutations in the exonuclease proof-reading domains of the DNA polymerases *POLD1* (codons 245-571) and *POLE* (codons 223-517) have been shown to predispose to the development of hypermutated microsatellite stable (MSS) sporadic colorectal cancer (16). Non-silent somatic mutations in *POLE* and *POLD1* were identified in 4 and 7 of the hypermutator cancers, respectively (Fig. 2; Supplementary Table S2A). Two shared somatic *POLE* mutations (P286R and F348S) affected the exonuclease proof-reading domain in cases 15G1 and 15G2, which were two adjacent cancers (adenocarcinoma and squamous cell carcinoma) in the same patient (Fig. 1A; Fig. 2).

No frameshift or non-silent somatic *POLE* or *POLD1* point mutations were identified in the non-hypermutator cancers.

The mechanism of hypermutation in case 6J remains undefined. Case 6J was an adenocarcinoma from a patient with a 20 year history of colonic Crohn's disease and autoimmune liver disease for which she received potent immunosuppression (prednisolone, tacrolimus and azathioprine) throughout this period. This patient had 4 basal and 2 squamous cell carcinomas of skin removed but did not have any frameshift or non-silent somatic or germline point mutations in *PTCH1*, excluding Gorlin syndrome as a contributing factor to the underlying cancer predisposition (17). In addition, germline mutations that could explain the hypermutator phenotype were not found in other DNA repair genes including those involved in nucleotide and base excision. It is possible that the continuous, prolonged immunosuppression in this patient has altered the immune cancer response resulting in both colonic and multiple skin cancers.

Mutational Signatures in IBD-CRC

It has been demonstrated that different mutational processes in cancers generate specific patterns of mutation, or 'signatures', with 30 distinct signatures identified thus far by Alexandrov *et al*. (18). The overall mutational spectrum our IBD-CRC cohort was very similar to the spectra derived from cohorts of sporadic CRCs from Giannakis *et al*. (14) (*cosine* similarity=.87) and from TCGA Data Portal (*cosine* similarity=.91) (Fig. 3A; Supplementary Materials and Methods). Six distinct signatures, designated A-F, were extracted from the catalogue of IBD-CRC somatic mutations (Fig. 3B) and corresponded well to Alexandrov Signatures 10, 1, 13/2, 17, 6 and 5, respectively, with *cosine* similarities ranging from .82 to .97 (Fig. 2 lower panel; Supplementary Table S3). Some of the signatures have been associated with specific mutational mechanisms (18). Of particular note, are IBD-CRC Signature A (Alexandrov Signature 10/*POLE*) that was predominant in the hypermutator cancers

with DNA *POLE* mutations (15G1 and 15G2); and IBD-CRC Signature E (Alexandrov Signature 6/dMMR and MSI) that was predominant in the 7 hypermutator cancers which had loss of expression of the MLH1 protein (Fig. 2). IBD-CRC Signature C (Alexandrov Signature 13 and 2/*AID/APOBEC*) is the major contributor in case 21M, however as we our analysis is limited to the exome regions only, we have no evidence of *AID/APOBEC* activation in this case. Transcriptional upregulation of *APOBEC3B* is commonly found in bladder, cervical, lung, head/neck and breast cancers with kataegis (18), however, we did not find evidence of kataegis in case 21M (Supplementary Fig. S4). It has been hypothesised that one cause of *APOBEC3B* upregulation may be infection by viruses, including Human Papillomavirus (HPV) (19). We did not find HPV DNA associated with case 21M, however, since whole-exome sequencing was used in this study, any viral sequence present would be excluded unless it has integrated into the genome in the targeted sequences.

Although Alexandrov signatures 2 and 13 have not yet been identified in CRC, an *APOBEC* mutation pattern has been found, using different methods, in a variety of cancer types including whole-genome sequenced sporadic CRCs (20). Signatures 1, 5, 6 and 10 have previously been found in sporadic colorectal cancers (18) and signature 17 has been identified in MSS CRC (21). Taken together, these results indicate that the mutational mechanisms in IBD-CRC and sporadic CRC are similar, although a much larger IBD-CRC cohort and whole-genome sequencing may reveal additional or novel signatures.

Recently, a signature attributed to the persistence of 8-oxoguanine G>T/C>A mismatches due to biallelic inactivation of *MUTYH* has been identified in *MUTYH*-associated polyposis (MAP) CRCs (22). Unsurprisingly, we did not find this signature

in our IBD-CRCs as our cases did not have a history of MAP nor any germline non-silent or splice site point or InDel mutations in *MUTYH* other than common SNPs. We did not discover a novel inflammation-associated mutational signature in our cohort. Similarly, others have not found an inflammation-signature common to chronic inflammatory-associated gastrointestinal cancers (such as Barrett's oesaphagus) (23). Rather, epithelial regeneration and the subsequent cumulative effect of chronic inflammation-associated damage appears to be a major mechanism of promoting carcinogenesis in IBD (24).

Driver Genes in IBD-CRC

To identify driver genes in non-hypermutator IBD-CRC, we used MutSigCV (9) to identify significantly mutated genes, and dNdScv to find genes under positive selection in cancer (10) (Methods and Supplementary Materials and Methods). In our 24 cases, *TP53*, *PIK3CA*, *APC*, and *KRAS* were identified as driver genes (FDR-adjusted *P*-value, or $q<.10$) (Fig. 4A), all of which are also driver genes in sporadic CRC (25). As observed in sporadic CRC (25), significantly more non-hypermutator cancers had non-silent somatic mutations in *TP53* than the hypermutator cancers (79% vs. 33%; Two-tailed Fisher's exact test, $P=.03$) (Table 2).

In the hypermutator cases, no driver genes were identified using MutSigCV (9) or MSMutsig (11), due to small sample size. When using dNdScv (10), followed by restricted hypothesis testing with a list of known cancer genes from the Cancer Gene Census (version 81) (12), *KRAS* ($q=.03$) was identified as a driver gene in the hypermutator cohort, in which 3 of the 4 cases with *KRAS* mutations are p.G12D (Fig. 4A). Importantly, this analysis also identified hotspot InDel mutations in *RPL22* (p.K15fs; $q=.001$), *ACVR2A* (p.K437fs; $q=.03$) and *RNF43* (p.G659fs; $q=.1$) (Fig. 4A; Supplementary Materials and Methods). The same *ACVR2A* and *RNF43* mutations

have previously been identified as driver events in a much larger cohort of sporadic CRC (11), and while they may represent driver events in IBD-CRC, they could also be genomic sites that are particularly prone to mutation. A much larger cohort of MSI IBC-CRCs would be required to confirm these as driver events. As reported for sporadic CRCs (26), we find enrichment of *RNF43* mutations in MSI cases, and mutual exclusivity of *RNF43* truncating mutations (which includes frameshift InDels, nonsense mutations and splice site mutations) with *APC* truncating mutations in our IBD-CRC cohort (Fig. 4A).

Several of the genes we identified are clinically actionable. For example, tumours with *RNF43* mutations are sensitive to porcupine inhibitors (27). In the same way *RPL22* mutant cancers may be susceptible to MDM2-p53 pathway compounds (28). Notably, the same *RPL22* p.K15fs hotspot mutation we identified in our IBD-CRC cohort has been frequently found in endometrial, gastric and colorectal cancers (28).

Somatic *IDH1* R132 Mutations

A potentially targetable *IDH1* hotspot mutation at R132, which was mutated in 11% (5/47) of IBD-CRCs and only in CD-associated CRCs, was previously reported by Yaeger *et al*. (29). In our cohort, one *IDH1* R132C mutation was present in UC-associated CRC case 32N3. The observed frequency here is not significantly different than that of Yaeger *et al*. (29) (*Chi* squared test, *P*=.21). We note, however, that the median 17-fold coverage (range 9-36-fold in the tumour samples) at this site was lower than the whole-exome median of 70-fold, as this mutation is only 20bp from an exon boundary where coverage tends to be lower in whole-exome sequencing studies.

Somatic Copy Number Alterations in IBD-CRC

As previously observed with sporadic CRC (30), somatic copy number alterations (SCNAs) in IBD-CRCs were significantly more prevalent (one-tailed Student's *t*-test *P*=.001) in the non-hypermutator cases compared with hypermutator cases (Supplementary Fig. S5; Supplementary Table S6).

Robles *et al*. (31) reported 15 focal copy number gains, 8 of which were observed in more than one tumour and Shivakumar *et al*. (32) reported 26 focal SCNAs in pooled IBD-associated dysplastic and carcinoma biopsy samples. Focal SCNAs common to our IBD-CRC cohort and these studies are gains in 12p13.33-12p13.31, 22q11.21, 4p16.3, 8q24.3, 10q26, 13q12, 20q11.23 and 20q13; however, none were common to all three.

We compared the frequency of predicted chromosome arm-level SCNAs to recurrent SCNAs found by Sheffer *et al*. (33), and to SCNAs identified by TCGA (25) (Fig. 4B; Supplementary Table S6). Our novel data demonstrate that the frequencies of specific arm-level SCNAs in IBD-CRC broadly concurs with those found in sporadic CRC (25, 33).

Germline Variants in IBD Patients

Colorectal cancer affects up to 2.5% of the patients who suffer from IBD (1), and germline variations conferring cancer susceptibility have been poorly described in this population of patients. We examined the germlines of our cohort for variants in genes implicated in susceptibility to colorectal cancer (Supplementary Tables S4 - S5). The *CHEK2* frameshift mutation at T367 (c.1100delC; rs555607708), in the germline of case 15N, is known to predispose to breast cancer, but association with colon cancer has not been resolved (34). The R242H mutation in *SDHB* (rs74315368) in case 3Q has been characterized in paragangliomas (35) but not, to

our knowledge, in colon cancer. The remaining germline variants we have reported in Supplementary Table S5 are in genes known to be associated with increased risk of CRC, however, they have unknown or mixed reported clinical significance and currently we are unable to clarify as to whether these variants confer a predisposition to colorectal cancer in the setting of chronic inflammation.

Neo-epitope and Immune Infiltrate Analysis

Non-synonymous somatic mutations in cancer can generate novel antigens, (neo-epitopes) that can be exploited in cancer immunotherapy. The number of predicted neo-epitopes is expected to be directly proportional to the mutational load of cancers (36). As expected, the IBD-CRC hypermutator cancers generated the largest number of HLA Class I neo-epitopes, and were significantly higher than the number found in non-hypermutators (one-tailed Student's *t*-test, *P*=.004) (Fig. 5). Similar to our study the neo-epitope load was higher in sporadic CRC with dMMR/MSI-high status, than in pMMR/MSS tumours, and notably, they responded favourably to PD1 blockade (14). IBD is characterized by a dysregulated immune response and many therapeutic modalities targeting inflammation are directed at cytokines. Analysis of the cytokine gene immune expression profile demonstrated that overall the cytokine gene expression profile was similar in the hypermutator IBD-CRCs and the non-hypermutator IBD-CRCs (Supplementary Fig. S6), although this analysis was limited by the very small size of tumour epithelium (with very little immune cell-containing stroma) sampled.

DISCUSSION

We have undertaken a comprehensive analysis of 34 IBD-CRCs from 31 patients using WES. Hypermutator cancers were observed in both our study and Robles *et al*.

(31) and the mutational rates were within range of the larger TCGA sporadic CRC cohort (25).

In sporadic and hereditary colorectal cancers, the hypermutator phenotype is most frequent in cancers of the right (proximal) colon (25) . In this study, we observed a strong association between CD-associated CRCs occurring in the proximal colon (Two-tailed *Fisher's* exact test, *P*=.03) and the hypermutator phenotype (Two-tailed Fisher's exact test, *P*=.05. Long standing, extensive colonic CD has been shown to be associated with proximal cancers (37) and we have clear histopathological evidence of pre-existing IBD at each tumour site, more than half of which were in the proximal colon. Similar to previous studies of IBD-CRC (38) a higher proportion of UC-associated CRCs occurred in the left colon (Fig. 2) when compared with CD-associated CRCs (Two-tailed Fisher's exact test, *P*=.03) .

Previous studies have reported a variable frequency of *MLH1* promoter hypermethylation and loss of expression of MLH1 in IBD-CRCs (39). The frequency of *MLH1* promoter methylation in our series may be associated with the more advanced age at time of IBD-CRC diagnosis, although it remains lower than that of sporadic MSI-high CRC with CpG (*MLH1*) hypermethylation (40). Lennerz *et al*. (39) have exclusively studied CRC complicating Crohn's colitis and described the median age of cancer diagnosis to be 58 years (range 34–77) which is also slightly higher than that conventionally reported for IBD-CRCs. In addition to our strict inclusion criteria, and similar to previous studies (39), the mutation analysis identified a single *BRAF* V600E mutation in one hypermutator IBD-CRC (Fig. 2) confirming that these cancers are unlikely to be sporadic in nature.

BRAF V600E mutations occur in the majority (>85%) of hypermethylated sporadic MSI high CRCs (25) and are therefore established in diagnostic algorithms to differentiate between sporadic and familial Lynch Syndrome cases of colorectal

cancer (41). Notably, we did not identify germline mutations in genes such as *MLH1*, *MSH2*, *APC*, and *MUTYH* that predispose to the development of CRC in our IBD-CRC cohort. Although the proportion of hypermutator cancers in our cohort was not significantly different than that in sporadic CRC, these differential mutations distinguish our hypermutated IBD-CRC cohort from hypermutated sporadic and familial forms of CRC.

The effect of dMMR/MSI on survival outcomes in IBD-CRC has not been reported. A recent meta analysis of 20 studies including 571,278 patients by Reynolds and colleagues (42) has reported that IBD-CRC does not affect the overall 5 year survival compared with sporadic CRC without any adjustment for molecular subtypes. In our series, the hypermutator IBD-CRCs (which included 2 MSS hypermutators 15G and 6J) had a significantly better survival compared with non-hypermutator IBD-CRCs. The increased survival of sporadic MSI cancers (43) is comparable to the data presented here for IBD-CRCs. The improved survival of early stage sporadic dMMR/MSI cancers is postulated to be mediated by (anti-) tumour infiltrating lymphocytes (TILs) in response to the neo-epitopes generated by the high mutational rate. Immune blockade therapies targeting immune checkpoints and enhancing the anti-TIL response are currently being used in various cancers including CRC (44). In our series, the hypermutator IBD-CRCs had a higher predicted neo-epitope load but we were unable to identify any obvious differences in the immune related gene expression profile. This is unsurprising as our samples had been enriched for cancer-containing cells and not the adjacent immune rich stromal compartment. Molecular phenotyping of all colorectal cancers for targeted therapy is a compelling reason for universal screening for dMMR . Of note, the United States Food and Drug Administration has approved Pembrolizumab (targeting the programmed cell death 1

receptor) for the treatment of all unresectable or metastatic MSI-H/dMMR tumours that have progressed after initial treatment (45). Immune checkpoint blockade has been associated with immune-mediated colitis and this is particularly relevant for patients with co-existing inflammatory bowel disease (46). Although our results require validation in a randomized prospective cohort, it is exciting to postulate that immune blockade therapies may be useful adjuncts in treating patients with dMMR (and/or hypermutator phenotype) associated IBD-CRC that have undergone a total colectomy as part of their treatment programme, negating the risk of immunotherapy mediated colitis. Importantly, immune therapies should be used with caution or avoided in IBD-CRC patients who have undergone a partial colectomy or those with Stage IV disease with an intact colon as it can aggravate the underlying colitis (47).

To date, we have not been able to use genetic or molecular markers to improve the detection or treatment of IBD-CRC. In 2009, a specialist committee in the United States of America recommended universal screening for Lynch syndrome in all newly diagnosed colorectal cancers (48) and this has not been adopted for several reasons including financial, technical, ethical and health economical limitations. In the United Kingdom, the National Institute for Health and Care Excellence (NICE) is advocating use of mismatch repair immunohistochemistry on all colorectal cancers (including IBD cases) to detect Lynch syndrome (41). A cost effective analysis limited to early onset colorectal cancers suggests that this would be economically viable assuming that all of the necessary subsequent health interventions are fully implemented to reduce cancer mortality and morbidity (49). Our data supports this recommendation to detect dMMR in IBD-CRC, but not necessarily to detect Lynch syndrome as we did not identify any known clinically significant pathogenic germline mutations in MMR genes in this IBD-CRC cohort. Whilst these important issues are resolved our data

can be used to support the rational for universal testing in IBD-CRCs to profile tumours for the use of targeted therapies, which is not routinely undertaken at all institutions,

The immunohistochemical analysis of *TP53* protein has not been widely accepted to aid in the discrimination between dysplastic and inflamed colonic epithelium. Prospective studies are required to determine whether analyzing for *TP53* abnormalities and loss of expression of MLH1 together, in colonic biopsies with potentially dysplastic epithelium, could aid in the evaluation of those at highest risk of developing CRC, similar to the molecular profiling used in gastric cancer (50).

Our retrospective study has inherent limitations and the small sample size may result in a reporting bias. Notwithstanding these issues, the power of next-generation sequencing technology has provided detailed information allowing clinically relevant statistical analyses to be undertaken. Future similar studies may discover many more similarities and differences between IBD-CRC and other types of CRC.

CONCLUSION

The categorisation of disease guides individualised therapeutic strategies and can predict the response to therapy and prognosis. Our analysis demonstrates that proximal IBD-CRCs have high mutational rates associated with defects in MMR and DNA *POLE* proof-reading function. This results in a predicted higher neo-epitope load, which could be exploited using immunotherapies. The hypermutator phenotype of this cohort is mostly associated with loss of MLH1 expression and this could be evaluated in colonic dysplastic lesions detected in IBD patients. The identification of driver genes in hypermutated IBD-CRCs can be used to develop therapeutic agents targeting the corresponding molecular pathways. In our series up to 90% of the

hypermutated IBD-CRCs have actionable mutations. These approaches would complement and individualise current surveillance and treatment programmes for IBD-CRC.

Acknowledgements

CpG Methylation primer sequences were generously provided by RH Masalmeh and D Sproul (University of Edinburgh).

Exome Sequencing:

All exome sequencing data is available from the European Genome-phenome Archive under accession number EGAS00001001129.

Author Contributions

SD acquired the ethical approval, provided the study samples, clinical data and analysis, created the cancer tissue micro-array, co-coordinated the study (with DJA and MJA), provided intellectual input on the project and interpretation of data, and drafted the manuscript.

KW analysed the sequencing data (including variant calling and validation, somatic copy number alterations, recurrently mutated genes and signatures of mutational processes), generated Figures and tables, and co-wrote the manuscript.

MFM conducted the *MLHI* promoter methylation analysis.

AO conducted the mismatch repair immunohistochemistry and associated analysis.

JH performed the DNA/RNA extractions and WES.

CJB reviewed and approved the IBD-CRC cases.

MLM and AJS conducted the HLA Neo-epitope predictions and critically reviewed the manuscript.

MR and RR performed the Nanostring analysis.

JS provided intellectual input and critically reviewed the manuscript.

DJA provided expert guidance, intellectual input on the project, coordinated and designed experiments with KW, JH, MM, AJS, MR and, RR and critically reviewed the manuscript.

MJA provided expert guidance, intellectual input on the project, reviewed the histology for the IBD-CRC cases, assessed the mismatch repair immunohistochemistry, designed experiments with MM and critically reviewed the manuscript.

REFERENCES

1. Sebastian, S., Hernandez, V., Myrelid, P., Kariv, R., Tsianos, E., Toruner, M., et al. Colorectal cancer in inflammatory bowel disease: results of the 3rd ECCO pathogenesis scientific workshop (I). J Crohns Colitis 2013, *8:* 5-18.
2. Sanduleanu, S. and Rutter, M. D. Interval colorectal cancers in inflammatory bowel disease: the grim statistics and true stories. Gastrointest Endosc Clin N Am 2014, *24:* 337-348.
3. RCoreTeam R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2015.
4. Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 2013, *31:* 213-219.
5. Saunders, C. T., Wong, W. S., Swamy, S., Becq, J., Murray, L. J., and Cheetham, R. K. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics 2012, *28:* 1811-1817.
6. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 2011, *27:* 2987-2993.
7. McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., et al. The Ensembl Variant Effect Predictor. Genome Biology 2016, *17:* 122.
8. Favero, F., Joshi, T., Marquard, A. M., Birkbak, N. J., Krzystanek, M., Li, Q., et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. Ann Oncol 2015, *26:* 64-70.
9. Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013, *499:* 214-218.
10. Martincorena, I. i., Raine, K. M., Gerstung, M., Dawson, K. J., Haase, K., Van Loo, P., et al. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell 2017, *171:* 1029-1041.e1021.

11. Maruvka, Y. E., Mouw, K. W., Karlic, R., Parasuraman, P., Kamburov, A., Polak, P., et al. Analysis of somatic microsatellite indels identifies driver events in human tumors. Nature Biotechnology 2017, *35:* 951.
12. Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., et al. A census of human cancer genes. Nature Reviews Cancer 2004, *4:* 177.
13. Gehring, J. S., Fischer, B., Lawrence, M., and Huber, W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. Bioinformatics 2015, *31:* 3673-3675.
14. Giannakis, M., Mu, XinmengÂ J., Shukla, SachetÂ A., Qian, ZhiÂ R., Cohen, O., Nishihara, R., et al. Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma. Cell Reports 2016, *15:* 857-865.
15. Lynch, H. T., Lynch, P. M., Lanspa, S. J., Snyder, C. L., Lynch, J. F., and Boland, C. R. Review of the Lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications. Clin Genet 2009, *76:* 1-18.
16. Heitzer, E. and Tomlinson, I. Replicative DNA polymerase mutations in cancer. Curr Opin Genet Dev 2014, *24:* 107-113.
17. Gailani, M. R., Bale, S. J., Leffell, D. J., DiGiovanna, J. J., Peck, G. L., Poliak, S., et al. Developmental defects in Gorlin syndrome related to a putative tumor suppressor gene on chromosome 9. Cell 1992, *69:* 111-117.
18. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., et al. Signatures of mutational processes in human cancer. Nature 2013, *500:* 415-421.
19. Vieira, V. C., Leonard, B., White, E. A., Starrett, G. J., Temiz, N. A., Lorenz, L. D., et al. Human Papillomavirus E6 Triggers Upregulation of the Antiviral and Cancer Genomic DNA Deaminase APOBEC3B. mBio 2014, *5:* e02234-02214.
20. Roberts, S. A., Lawrence, M. S., Klimczak, L. J., Grimm, S. A., Fargo, D., Stojanov, P., et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nat Genet 2013, *45:* 970-976.
21. Katainen, R., Dave, K., Pitkanen, E., Palin, K., Kivioja, T., Valimaki, N., et al. CTCF/cohesin-binding sites are frequently mutated in cancer. Nat Genet 2015, *47:* 818-821.
22. Sieber, O. M., Lipton, L., Crabtree, M., Heinimann, K., Fidalgo, P., Phillips, R. K. S., et al. Multiple Colorectal Adenomas, Classic Adenomatous Polyposis, and Germ-Line Mutations in MYH. New England Journal of Medicine 2003, *348:* 791-799.
23. Matsumoto, T., Shimizu, T., Takai, A., and Marusawa, H. Exploring the Mechanisms of Gastrointestinal Cancer Development Using Deep Sequencing Analysis. Cancers 2015, *7:* 1037-1051.
24. Choi, C. R., Bakir, I. A., Hart, A. L., and Graham, T. A. Clonal evolution of colorectal cancer in IBD. Nat Rev Gastroenterol Hepatol 2017.
25. The Cancer Genome Atlas, N. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012, *487:* 330-337.
26. Giannakis, M., Hodis, E., Jasmine Mu, X., Yamauchi, M., Rosenbluh, J., Cibulskis, K., et al. RNF43 is frequently mutated in colorectal and endometrial cancers. Nat Genet 2014, *46:* 1264-1266.
27. Koo, B.-K., van Es, J. H., van den Born, M., and Clevers, H. Porcupine inhibitor suppresses paracrine Wnt-driven growth of mutant neoplasia. Proceedings of the National Academy of Sciences 2015, *112:* 7548-7550.
28. Ferreira, A. M., Tuominen, I., van Dijk-Bos, K., Sanjabi, B., van der Sluis, T., van der Zee, A. G., et al. High frequency of RPL22 mutations in microsatellite-unstable colorectal and endometrial tumors. Hum Mutat 2014, *35:* 1442-1445.
29. Yaeger, R., Shah, M. A., Miller, V. A., Kelsen, J. R., Wang, K., Heins, Z. J., et al. Genomic Alterations Observed in Colitis-Associated Cancers Are Distinct From

Those Found in Sporadic Colorectal Cancers and Vary by Type of Inflammatory Bowel Disease. Gastroenterology 2016, *151:* 278-287 e276.
30. Wang, H., Liang, L., Fang, J. Y., and Xu, J. Somatic gene copy number alterations in colorectal cancer: new quest for cancer drivers and biomarkers. Oncogene 2016, *35:* 2011-2019.
31. Robles, A. I., Traverso, G., Zhang, M., Roberts, N. J., Khan, M. A., Joseph, C., et al. Whole-Exome Sequencing Analyses of Inflammatory Bowel Disease-Associated Colorectal Cancers. Gastroenterology 2016, *150:* 931-943.
32. Shivakumar, B. M., Chakrabarty, S., Rotti, H., Seenappa, V., Rao, L., Geetha, V., et al. Comparative analysis of copy number variations in ulcerative colitis associated and sporadic colorectal neoplasia. BMC Cancer 2016, *16:* 271.
33. Sheffer, M., Bacolod, M. D., Zuk, O., Giardina, S. F., Pincas, H., Barany, F., et al. Association of survival and disease progression with chromosomal instability: a genomic exploration of colorectal cancer. Proc Natl Acad Sci U S A 2009, *106:* 7131-7136.
34. Cybulski, C., Wokolorczyk, D., Kladny, J., Kurzwaski, G., Suchy, J., Grabowska, E., et al. Germline CHEK2 mutations and colorectal cancer risk: different effects of a missense and truncating mutations? European Journal of Human Genetics 2007, *15:* 237-241.
35. Young, A. L., Baysal, B. E., Deb, A., and Young, W. F., Jr. Familial malignant catecholamine-secreting paraganglioma with prolonged survival associated with mutation in the succinate dehydrogenase B gene. J Clin Endocrinol Metab 2002, *87:* 4101-4105.
36. Platten, M. and Offringa, R. Cancer immunotherapy: exploiting neoepitopes. Cell Res 2015, *25:* 887-888.
37. Bansal, P. and Sonnenberg, A. Risk factors of colorectal cancer in inflammatory bowel disease. Am J Gastroenterol 1996, *91:* 44-48.
38. Choi, P. M. and Zelig, M. P. Similarity of colorectal cancer in Crohn's disease and ulcerative colitis: implications for carcinogenesis and prevention. Gut 1994, *35:* 950-954.
39. Lennerz, J. K., van der Sloot, K. W. J., Le, L. P., Batten, J. M., Han, J. Y., Fan, K. C., et al. Colorectal cancer in Crohn's colitis is comparable to sporadic colorectal cancer. International Journal of Colorectal Disease 2016, *31:* 973-982.
40. Weisenberger, D. J., Siegmund, K. D., Campan, M., Young, J., Long, T. I., Faasse, M. A., et al. CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. Nat Genet 2006, *38:* 787-793.
41. https://www.nice.org.uk/guidance/dg27/documents/diagnostics-consultation-document Molecular testing strategies for Lynch syndrome in people with colorectal cancer. 2017.
42. Reynolds, I. S., Oâ€™Toole, A., Deasy, J., McNamara, D. A., and Burke, J. P. A meta-analysis of the clinicopathological characteristics and survival outcomes of inflammatory bowel disease associated colorectal cancer. International Journal of Colorectal Disease 2017, *32:* 443-451.
43. Sinicrope, F. A. and Sargent, D. J. Molecular Pathways: Microsatellite Instability in Colorectal Cancer: Prognostic, Predictive, and Therapeutic Implications. Clinical Cancer Research 2012, *18:* 1506-1512.
44. Le, D. T., Uram, J. N., Wang, H., Bartlett, B. R., Kemberling, H., Eyring, A. D., et al. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. N Engl J Med 2015, *372:* 2509-2520.

45.https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm560167.htm FDA approves first cancer treatment for any solid tumor with a specific genetic feature. 2017.
46. Michot, J. M., Bigenwald, C., Champiat, S., Collins, M., Carbonnel, F., Postel-Vinay, S., et al. Immune-related adverse events with immune checkpoint blockade: a comprehensive review. European Journal of Cancer 2016, *54:* 139-148.
47. Johnson, D. B., Sullivan, R. J., Ott, P. A., and et al. Ipilimumab therapy in patients with advanced melanoma and preexisting autoimmune disorders. JAMA Oncology 2016, *2:* 234-240.
48. Evaluation of Genomic Applications in Practice and Prevention Working, G. Recommendations from the EGAPP Working Group: genetic testing strategies in newly diagnosed individuals with colorectal cancer aimed at reducing morbidity and mortality from Lynch syndrome in relatives. Genetics in Medicine 2009, *11:* 35-41.
49. Snowsill, T., Huxley, N., Hoyle, M., Jones-Hughes, T., Coelho, H., Cooper, C., et al. A model-based assessment of the cost-utility of strategies to identify Lynch syndrome in early-onset colorectal cancer patients. BMC Cancer 2015, *15:* 313.
50. Cristescu R, Lee J, Nebozhyn M, Kim K, and J, T. Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. Nature Medicine 2015, *21:* 449–456.

Table 1 Mutational rates in IBD-CRC and sporadic CRC.

TUMOUR PHENOTYPE	TCGA(25)	DIN	ROBLES(31)
Hypermutators	>12 (728)	32.6-171.3 (56.2)	18.19-22.2 (N/A)
Non-hypermutators	<8.24 (58); Non-silent mutations only	2.1-7.0 (3.1)	0.3-5.1 (1.3)

Mutational rates (mutations/Mb) from the whole-exome sequence analysis of TCGA sporadic CRC cohort (25) and IBD-CRCs from this study and Robles *et al*. (31) in the hypermutator and non-hypermutator cancers. Numbers in parentheses are median mutation rates. Two hypermutators were present in the Robles cohort. The rates provided include total mutations with the exception of the median mutation rate of non-hypermutator cancers in the TCGA cohort.

Table 2 Frequency of *TP53, KRAS* and *APC* mutations in IBD-CRC and sporadic CRC.

GENE	TCGA HM(25)	TCGA NHM(25)	DIN HM	DIN NHM	ROBLES NHM(31)
TP53	20%*	60%*	33%*	79%*	63%
KRAS	30%	43%	44%	25%	20%
APC	51%*	81%*	33%	29%	13%
PIK3CA	13%	23%	56%	29%	10%

Mutational frequencies from the whole-exome sequence analysis of TCGA sporadic CRC cohort (25) and IBD-CRCs from this study and Robles *et al*. (31) in hypermutator (HM) and non-hypermutator (NHM) cancers.
*Significant differences between HM and NHM CRCs.

Figure Legends

Figure 1: Somatic mutational rates and survival analysis of IBD-CRC.

(**A**). Mutational frequency in each of the 34 IBD-CRCs ordered by overall mutation rate. There is a clear separation between the 10 hypermutator cancers and the 24 non-hypermutator cancers. With the exception of 15G1, 15G2, and 6J, the hypermutator cancers showed elevated mutation rates of both SNVs and InDels. No InDels were found in the exome of 21M. (**B**) Kaplan-Meier plot of overall survival stratified by cancer mutator phenotype. Patients with hypermutators cancers had

increased survival compared with patients with non-hypermutator cancers (log rank test, *P*=.04).

Figure 2: Clinical and genetic characteristics of the IBD-CRC.

Top panel: Mutational rate of each cancer case ordered by tumour site and mutation rate. The red line segregates the proximal and distal bowel cancers. Second panel: Tumour site, Dukes' stage, cancer type, underlying IBD, immunohistochemical testing of MLH1 (MLH1 IHC) and promoter methylation status of *MLH1* (*MLH1* Pr-M). Third panel: selected genes and somatic nonsense, non-silent and InDel mutations. The colours indicate the predicted effect of the mutation on the protein sequence, and the number indicates the number of mutations with the same predicted effect. Note that in cases where a gene has multiple mutations with different predicted effects, only the effect type with the highest priority will be shown. Further details are in the Supplementary Methods and Materials. The complete list of mutations and their effects is in Supplementary Tables S2A - S2B. Lower panel: Contribution of signatures of mutational processes (A-F) in each cancer case. Each signature A-F corresponds to a signature identified by Alexandrov *et al*. (Signatures 10, 1, 13/2, 17, 6 and 5, respectively) with *cosine* similarities ranging from 0.82 to 0.97 (Supplementary Table S3).

Figure 3: Signatures of mutational processes in IBD-CRC.

(**A**). The somatic mutation spectra of the IBD-CRC and sporadic CRC datasets are represented by each of the six possible substitution types, C>A, C>G, C>T, T>A, T>C, T>G (including their reverse complements) in the context of their immediate 5' and 3' flanking bases, giving 96 possible motif combinations. For each substitution type, the 16 motifs are listed in order from left to right by the 5' flanking base (A, C, G,

then T), then by the 3' flanking base (A, C, G, then T). The IBD-CRC panel shows the mutation spectrum for 34 IBD-associated colorectal cancers; the CRC (TGCA) panel shows the mutation spectrum for 115 colon adenocarcinomas and 267 rectum adenocarcinomas from downloaded from TCGA Data Portal, and the CRC (NHS/HPFS) panel show the mutation spectrum for a cohort of 619 CRCs from Giannakis *et al*. (14). (**B**) Six distinct mutation signatures (A-F) were extracted from IBD-CRCs and on comparison with the Alexandrov Signatures 1 to 30 (33) using cosine similarity (Supplementary Table S3), the correspondence is: Signature A and Signature 10 (*POLE* mutations); Signature B and Signature 1 (age/spontaneous deamination of 5-methylcyotosine); Signature C and Signature 13/2 (*AID/APOBEC* activation); Signature D and Signature 17 (unknown aetiology); Signature E and Signature 6 (mismatch repair deficiency and microsatellite instability); Signature F and Signature 5 (unknown aetiology; found in all cancer types).

Figure 4: Hotspot microsatellite InDels, driver genes and recurrent somatic copy number alterations in IBD-CRC genomes.

(**A**). Hotspot InDels and driver genes in IBD-CRC. Somatic SNVs and InDels were used to identify hotspot InDels and driver genes in hypermutator (*n*=9) and non-hypermutator cancers (*n*=24). Because hypermutator case 15G1 and 15G2 were related, SNV and InDels were merged for this analysis. (**B**) Chromosome arm-level somatic copy number alterations. Shown are predicted SCNAs that cover at least 90% of a chromosome arm (left panel), and the frequency of the SCNAs in IBD-CRC (this study), and in sporadic CRC studies by Sheffer *et al*. (33) and TCGA (25) (right panel). Chromosome arms are listed in descending order of frequency in IBD-CRC. Grey boxes in the frequency columns indicate unknown values. IBD-CRC cases are ordered left to right by overall SNV and InDel mutation rate. Cases to the left of the dashed line are hypermutator cancers.

Figure 5: Predicted HLA Class I neo-epitopes in IBD-CRCs.

The cancers with the highest mutational rates (hypermutators) generated the largest number of predicted HLA Class I neo-epitopes**.**

SUPPLEMENTARY MATERIALS AND METHODS

Tumour Tissue Microarray

Formalin fixed paraffin embedded (FFPE) tumour and uninvolved lymph node blocks were sectioned and H&E stained to select suitable areas for macro-dissection. Six mm cores were used for creation of the tumour microarray.

DNA/RNA Extractions, Exome Sequencing and Data Processing

Nucleic acids were extracted using Qiagen Allprep FFPE DNA extraction kits, and DNA/RNA was quantified using Agilent geneChips.

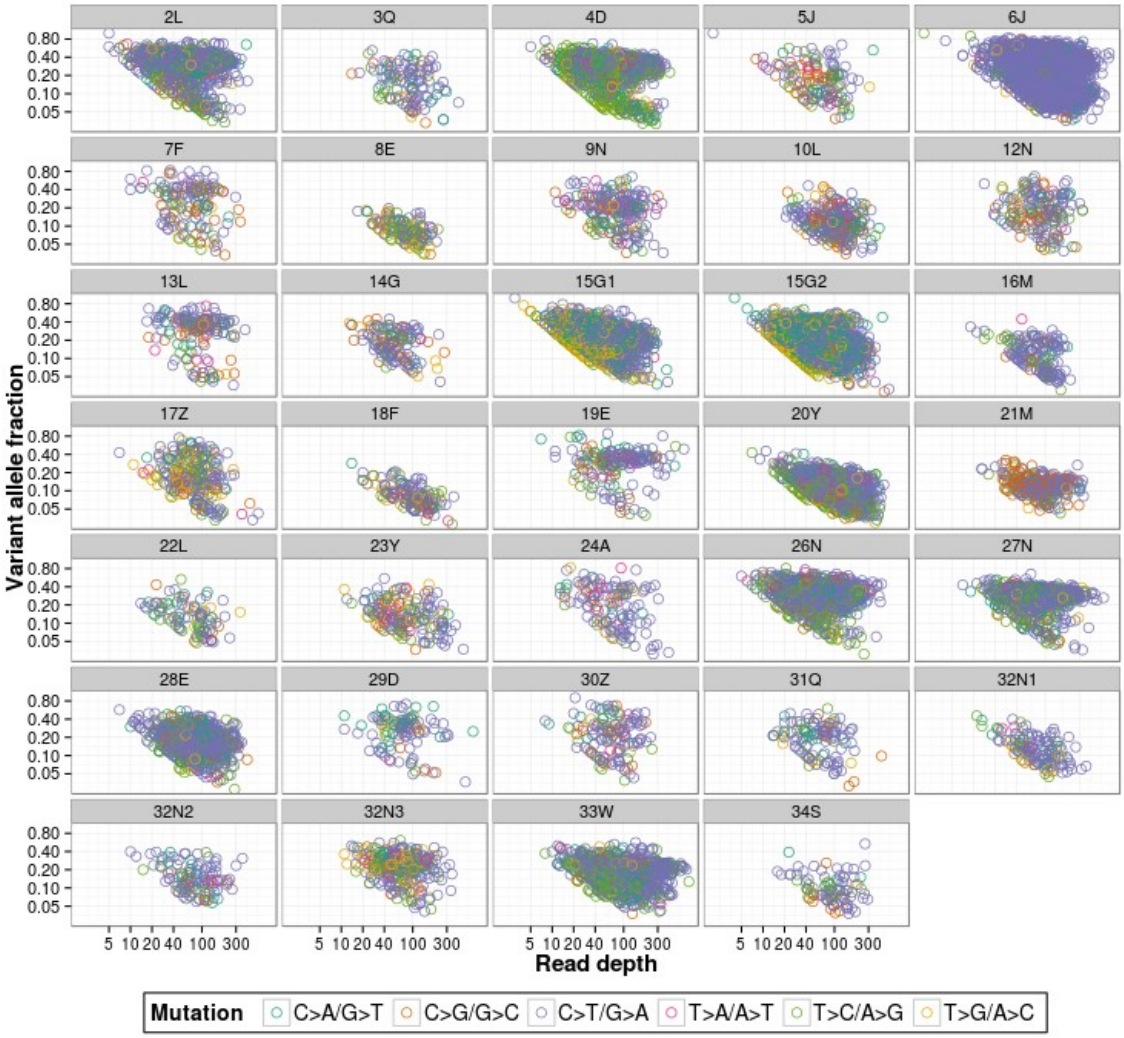
For sequencing library generation, genomic DNA was fragmented for each tumour and normal sample. Enrichment for exome regions was performed using the Agilent SureSelectXT Human All Exon v5 platform following the manufacturer's protocol. Six PCR cycles were performed for library preparation, and two PCR cycles were performed at the hybridization stage. Captured fragments were indexed and sequenced using the Illumina HiSeq 2000 platform at the Wellcome Trust Sanger Institute. Raw paired-end sequencing reads (75bp) were aligned to the reference genome hs37d5 (as used in Phase 3 of the 1000 Genome Project (1)), which includes the GRCh37 primary assembly and additional human contigs and viral sequences that reduce the number of reads erroneously mapped to the primary assembly. Alignments were performed using the 'aln' and 'sampe' algorithms of Burrow-Wheeler Aligner (BWA) software package (version 0.5.10) (2) to produce binary Sequence Alignment/Map (BAM) files (2). The option used for BWA 'aln' was '-q 15' and default parameters were used for 'sampe'. Base quality score recalibration (BSQR) and read realignment around known common InDels was performed using

tools from the Genome Analysis Tool Kit (GATK) (3) with default parameters. SNPs from dbSNP build 135 (4) were used as known sites for BQSR, and known InDels as used by the 1000 Genomes Project (1) were used for realignment around InDels. All files were downloaded from the 1000 Genomes Project FTP site: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_mapping_resources/. Per sample, between 0.004% to 11% (median 1%) of reads were unmapped, and 16%-31% (median 22%) of reads were PCR duplicates. The average Phred scaled mapping quality ranged from 24-34 (median 30). PCR duplicates were flagged with Biobambam (version 0.0.188) function 'bammarkduplicates' (5), and removed from BAM files. Secondary read alignments were also removed, as were reads that failed Illumina chastity (purity) filtering. Upon visual inspection of read alignments, we discovered apparent chimeric PCR duplicates comprised of one read mate mapped to distant repeat elements, and these read pairs were also excluded from the BAM files for downstream analyses. Among all samples, the portion of reads mapping to exome bait regions plus 100 bp flanking sequence was 53%-77% (median 71%). Variant calling was not performed on reads aligning outside of these regions. Variants called at the *BAP1* locus resulting from a contaminating plasmid were removed from all downstream analyses.

All raw sequencing data is available from the European Genome-phenome Archive under accession number EGAS00001001129.

Assessment of Deamination Artefacts

To assess deamination artefacts, for each sample we generated plots of read depth versus variant allele fraction (VAF) of all somatic point mutations calls (see plot below).



Artefacts from 5mC deamination (C>T/G>A) are typically observed at ~10% VAF and lower, at a wide range of read depths. We did not visualize an excess of C>T/G>A at these levels, however, as many of the samples appeared to have low cellularity (non-C>T/G>A mutations with 10-20% VAF) it was visually difficult to estimate the level of 5mC deamination artefacts. This also made filtering of mutation calls using a VAF cutoff (in addition to those already applied by the variant calling software) not practical since many true somatic mutations would be lost. To further evaluate deamination artefacts and false positives, we performed validation experiments, described below and in the “Validation of Somatic Variants” section.

From validation experiments using the Sequenom MassARRAY iPLEX genotyping platform, the overall false positive rate for the point mutations was estimated to be 15% (see below). Approximately 68% (40/59) of the false positive point mutation calls were C>T or A>G. This makes up 16% (40/254) of the total C>T or A>G mutations with successful base calls from Sequenom MassARRAY iPLEX genotyping. The majority of these are likely artefacts from 5mC deamination. As the overall false positive rate is 15%, there will be a small effect on calculation of mutation rates and neo-epitope predictions; however, the effect is not large enough to change the main conclusions regarding mutator status and neo-epitope load. Artefacts resulting from 5mC deamination occur in a specific context but in random genomic locations, and at low levels, would have little or no effect on the discovery of driver genes. There is a known mutational signature associated with C>T/A>G substitutions originating from 5mC deamination (Alexandrov signature 1), and these would be classified as such in the mutational signature analysis. Other false positive substitution types would not produce any mutation signature unless the false positives were a result of a context-specific artefact. We also visually inspected the sequencing read alignments at predicted somatic mutation sites in the genes shown in Fig. 2/Supplementary Table 2A and 2B, Supplementary Table S5, and significantly mutated genes to check for false positive substitutions that may have resulted from reads that have potentially been mis-mapped.

Somatic Mutation Calling With MuTect

Somatic site mutations were identified from comparison of tumour and matched normal BAM files using MuTect (v1.1.7) (6). Default parameters were used, except in the following:

--max_alt_allele_in_normal_fraction 0.1

--max_alt_alleles_in_normal_count 5

--pir_median_threshold 12

--pir_mad_threshold 1.5

--heavily_clipped_read_fraction 0.25

A dbSNP file (build 141) (7) and COSMIC file (v73; common SNPs removed) (8) were used as input to MuTect. Further filtering of calls is described below.

Somatic InDel Calling With Strelka

Strelka (v1.0.14) (9) was used to identify somatic InDels. The configuration file available in the Strelka package for reads aligned with BWA was used. Default parameters were used, with the exception of the parameters below:

isSkipDepthFilters = 1

indelMaxRefRepeat = 100

indelMaxIntHpolLength = 50

sindelNoise = 0.0000005

minTier1Mapq = 15

sindelQuality_LowerBound = 25

isWriteRealignedBam = 1

Additional arguments were provided:

'-min-candidate-indel-reads 4 -min-small-candidate-indel-read-frac 0.05 -min-candidate-indel-read-frac 0.05'

Further filtering of calls is described below.

Filtering SNVs and InDels

SNVs and InDels falling within the bait regions defined by the Agilent SureSelectXT Human All Exome v5 bait regions plus 100bp flank were included in analyses. Variants were compared to those found in the ExAc database (10) and removed if present with a population allele frequency >=1%. Variants determined to be false positives from Sequenom MassARRAY iPLEX genotyping were removed. Read alignments for genes of interest (Fig. 2 in the main text) were inspected using the Integrative Genomics Viewer (IGV) (11) to verify the presence of somatic mutations, and to compare 15G1 and 15G2 for calls that may have been missed.

In counting and analysis of mutations in the exomes, all mutations falling within exons and affecting splice acceptor and splice donor sites were (12)included.

Predicting Variant Consequences With Variant Effect Predictor (VEP)

The Ensembl Variant Effect Predictor (VEP) (12) software was used to predict the effect of somatic and germline variants on protein sequences. Gene models from Ensembl (13) release 81 which is based on the reference genome version GRCh37 (14) were used. Variants in genes with the Ensembl biotype designation 'protein coding' and those affecting protein coding regions or essential splice sites in targeted exome regions were retained for downstream analyses, and only the effects on canonical transcripts were considered. For genes shown in Fig. 2, if multiple mutations were observed in the same gene, and the predicted effects of the mutations on the protein sequence were different, then only the highest priorty effect was shown in Fig. 2. The priority, from highest to lowest, was "splice acceptor variant", "stop gained", "framehshift variant", "inframe insertion", "inframe deletion", and "missense variant". This order of priority is based on the estimated severity of the effects as determined by Ensembl (http://www.ensembl.org/info/genome/variation/predicted_data.html). The full list of

mutations and consequences for genes in Fig. 2 is in Supplementary Tables S2A - S2B.

Validation of Somatic Variants

A random selection of 529 somatic variant sites (379 SNVs, 110 small deletions and 40 small insertions) affecting coding regions or essential splice sites, plus 50 non-silent mutations in the genes *APC*, *KRAS*, *MLH1*, *MSH2*, *MSH3*, *MSH6*, *POLE*, *POLD1*, *TP53*, *MUTYH*, *APOBEC3A* and *APOBEC3B* (41 SNVs, 8 small deletions, 1 small insertion) were subjected to validation using the Sequenom MassARRAY iPLEX genotyping platform 15. Eight of the 579 variant sites selected were also called in at least one other tumour sample; including these recurrent variants, a total of 595 variant sites were available to estimate the false positive rates. Genotyping assays were performed across all tumours and matched normal samples. Fourty-nine validation sites either failed the Sequenom assay or had an 'N' base call in the sample with the predicted variant base (or its matched normal), leaving 546 assays to estimate the false positive rates which were 15% (59/402) for SNVs, 11% (12/109) for small deletions, and 23% (8/35) for small insertions. Of the mutations in the genes of interest in Fig. 2, for 41% we were unable to perform validation, primarily due to lack of DNA. The remainder of the mutations in Fig. 2 have been verified as somatic mutations. Eight mutations in these genes were false positives and not included in Fig. 2. We used PCR amplification and Sanger sequencing to confirm the presence or absence of each hotspot InDel in *ACVR2A*, *RPL22*, and *RNF43* (Fig. 4A) in 8 hypermutator tumour samples and their matched normal samples to confirm the somatic status of the InDels. We did not have sufficient DNA for 2 hypermutator samples, 20Y and 15G2, and thus these samples were excluded. In the remaining samples, the somatic hotspot InDels in these 3 genes were confirmed, with the exception of the *RPL22* p.K15fs mutation which we could not confirm due to low

quality sequencing results. Examples of chromatograms generated from Sanger sequencing validation are shown in Supplementary Fig. S7. All chromatograms were visually inspected for the presence or absence of the expected InDel, and confirmed using the web application Indigo (https://gear.embl.de/indigo). We also confirmed the presence or absence of the *DOCK3* p.T1850fs mutation in the hypermutator cases (Supplementary Table S7), which was identified as a driver event in a much larger cohort of sporadic CRC (15) but failed to reach significance in our analysis. All read alignments for mutations in Fig. 2 and Fig. 4A were also visually inspected using the Integrated Genomics Viewer (11). A complete list of somatic mutations is available in Supplementary Table S7 - S8, and those that have been validated using the Sequenom MassARRAY platform and Sanger sequencing are indicated.

The region of interest in the *RPL22*, *RNF43*, *ACVR2A*, and *DOCK3* genes was amplified from genomic DNA using Thermo Scientific Thermo-Start *Taq* DNA Polymerase (following manufacturer's instructions) and the oligo pairs below. The expected product sizes were 301, 282, 284, and 276bp, respectively. The resulting amplified products were sequenced by Sanger sequencing (Eurofins) using the same oligos.

RPL22 Forward	5' CCCCAGTAGGTTTTCTCAACA 3'
RPL22 Reverse	5' AGCACACTTCCGTTAGTTTTG 3'
RNF43 Forward	5' GTCCACAGATCAAGGGGTGT 3'
RNF43 Reverse	5' GGCTCTCTAACCCACAGTGC 3'
ACVR2A Forward	5' CCAGTTTGAAAGTCAGGAGGA 3'
ACVR2A Reverse	5' TGCAGAAGAAAGAGAAATGTGC 3'
DOCK3 Forward	5' CCCAGGGGACTCTTCTCAT 3'
DOCK3 Reverse	5' TTCCTCCTGTAGACCCCAAG 3'

Annotation of Variants With ClinVar Clinical Significance

A ClinVar database (16) VCF file (release date 2016-05-31) was downloaded from the ClinVar FTP site (17). All somatic mutations were annotated with ClinVar clinical significance. To identify potentially clinically relevant variants in the germline exomes, variants identified in the normal lymph node samples were also compared to variants in the files 'common_and_clinical.vcf.gz' and 'common_no_known_medical_impact.vcf.gz' (release date 2016-05-31), also downloaded from the ClinVar FTP site. These represent variants that are common in the general population but are clinically relevant, and common variants that are not known to have medical impact, respectively. Filtering of germline variants using this information is described below.

Variant Calling in Germline Exomes

For each of the normal lymph node samples, SAMtools (v1.3) (18) and bcftools (v1.3) (19) were used to identify variants relative to the human reference genome GRCh37, and bcftools 'norm' was used to left-align all InDel calls and split mutliallelic sites into multiple rows. Soft-filtering of variant calls was performed using bcftools 'filter' and only calls annotated as 'PASS' were considered for downstream analyses. The following parameters were used:

SAMtools mpileup -C50 -pm4 -F0.2 -d1000 -ug -t DP,DPR,DV,DP4,INFO/DPR,SP

bcftools call -vm

bcftools norm -m - –D –T baits_w100.bed

bcftools filter -m+ -sLowQual -e\"%QUAL<=10\" -g3 -G5 -Ov –T baits_w100.bed

where the file baits_w100.bed is a BED file containing genomic regions defined by the Agilent SureSelectXT Human All Exome v5 baits with 100bp flank on each side.

Variant consequences were predicted using the Variant Effect Predictor as described above. We identified germline variants in genes implicated in familial CRC (Supplementary Table S4) (20) and searched for these variants in the ClinVar database (release 2016-05-31) (16) to determine their clinical significance. We removed variants that are found in dbSNP build 147 (21) which were common SNPs with no known medical impact (see above) and variants found in the ClinVar database and classified with a clinical significance of 'benign' or 'likely benign'. Variants found in the ClinVar database classified as 'pathogenic', 'unknown' or 'untested', common SNPs with known medical impact (see above), or not found in the ClinVar or dbSNP databases were considered variants of interest for Supplementary Table S5 if the variant was absent from the ExAc database (10) or present with an allele frequency less than 0.01.

Identification of Somatic Copy Number Alterations (SCNA) with Sequenza

The Sequenza software package (version 2.1.2) (22) was used to estimate tumour cellularity and ploidy, and to obtain allele-specific copy number profiles (Supplementary Table S6). For data preprocessing, SAMtools (v1.3) (18) was used to generate 'pileup' files for each tumour an normal sample BAM file, with '-Q 20' to skip bases with a quality score less than 20. The 'pileup2seqz' function in the sequenza-utils.py script was used to generate input files for Sequenza (v2.1.2) (22). All best-fit and alternative solutions for estimates of cellularity and ploidy were inspected manually using the contour plots, model fit and alternative fit plots, and chromosome copy number plots that are output by Sequenza. Samples with potentially noisy data were identified as those with a high estimated ploidy. Those with a weaving pattern of alternating copy number gain and loss as visualized in a genome-wide copy number plot can potentially be due to noise, or cases with a high

number of true SCNAs. In cases with ploidy greater than three, an alternative solution with ploidy closest to 2 was chosen. However, in these cases, some of the copy number alteration predictions may be inaccurate.

For each IBD-CRC sample, SCNAs were compared to recurrent SCNAs in sporadic CRC identified in Sheffer *et al*. (23) and SCNAs found in the TCGA cohort (24). We defined chromosome arm-level SCNAs as those with at least 90% of the chromosome arm deleted or amplified. All 18 broad SCNAs from Sheffer *et al*. (23) were observed in at least 1 IBD-CRC case, and many IBD-CRC cases have more than one recurrent Sheffer SCNA. Additionally, many of the chromosome arm-level SCNAs were observed at similar frequencies in IBD-CRC and both sporadic CRC data sets, providing confidence that the patterns of gains and losses observed in the genome-wide copy number plots reflect true changes rather than noise (See also Supplementary Table S6).

Identification of Driver Genes

For somatic mutations in the non-hypermutator cohort, we applied the software MutSigCV (version 1.4) (25) to find genes mutated more often than expected among the IBD-CRC tumour samples. Both somatic point mutations and InDels were used for this analysis. The required coverage file and covariates file were downloaded from the MutSigCV website (26) and the workflow described on the website was applied. Mutation Annotation Format (MAF) files where generated from the Variant Call Format (VCF) files produced by MuTect, and annotated by VEP (see above), using in-house Perl scripts. A FDR-adjusted *P*-value (*q*-value) cutoff of .1 was applied to select significant genes. All sequencing read alignments from SNV and InDel mutations in driver genes were visually inspected in the cancer and matched normal lymph node using the Integrated Genomics Viewer (11). For the non-hypermutators,

mutations occurring in exons and 50 bp surrounding exons were included to increase the number of mutations available for MutSigCV. We also used the dNdScv algorithm (27), which looks for positive selection in cancer using both SNVs and InDels in protein coding and splice site regions. After running dNdScv with default settings, we performed restricted hypothesis testing using a set of known cancer genes from the Cancer Gene Census (CGC) (version 81) (28). Both MutSigCV and dNdScv identified *TP53*, *PIK3CA*, *APC*, and *KRAS* as driver genes (q<.1) in the non-hypermutator cohort.

For the hypermutator cohort, MutSigCV was run as described above, except InDels falling inside of microsatellites (see below) were removed prior to running MutSigCV. In addition, because cases 15G1 and 15G2 appear to have originated from a common precursor clone, the union of the mutation catalogs for these samples was used to avoid counting shared mutations twice. No significant genes were identified, likely due to lack of power (n=9). We then obtained custom scripts, required input files, and a modified version of MSMutSig from the author (15), which enabled us to run MSMutSig to identify significantly mutated genes using microsatellite InDels from our 7 MSI samples. MSMutSig was run with default settings, and with the assumption that all loci were covered in all samples. Again, no significant results were obtained, due to lack of power. We then applied dNdScv, which is able to analyze both the SNV and InDel mutations together in the hypermutator cohort (27). By default, dNdScv limits the number of mutations per gene per sample to 3, and the maximum coding mutations per sample to 3000. This excluded 15G1/2 and 4D from the analysis. Two InDel models are available in dNdScv; one model considers the total number of InDel mutations per gene, and the other model considers unique InDel sites per gene (the “unique-sites model”). The unique-sites model was shown to work well to identify true drivers in MSI tumours, while using the total number of InDels

finds recurrent InDel sites which may or may not be true driver mutations (27). When dNdSV was applied to our IBD-CRC hypermutator cases using the “unique-sites” model, and restricted hypothesis testing was performed as described above, only *KRAS* was identified as a driver gene (q=.08). We then used the second model for InDels, which allows dNdScv to consider the total number of InDels, and performed restricted hypothesis testing as described above. Four genes, *ACVR2A (*q=.03), *KRAS* (q=.03), *RNF43* (q=.1), and *RPL22* (q=.00006) were identified. By allowing a maximum FDR of 0.1, we expect < 1 false positive in this gene list. The hotspot InDels in *ACVR2A*, *RNF43* and *RPL22* (Fig. 4A) were validated using PCR amplification and Sanger sequencing in all hypermutator samples except 15G2 and 20Y which were depleted of DNA. See “Validation of Somatic Variants” above.

Microsatellite repeats, defined as repeats 1-6bp in size with a minimum of 5 repeat units, were identified in the reference genome using Phobos (29) the parameters used are listed below:

-U 6

--minPerfection 85

–minScore 4

–preferShorterRepeats

–dontRemoveMostlyOverlapping

Repeats with 5 or more repeat units were extracted from the output and somatic InDels falling inside any of these regions were removed from the HM MSI samples (all HMs except for 15G1/2 and 6J) prior to analysis with MutSigCV.

Mutational Spectra and Extraction of Mutational Signatures Using SomaticSignatures

Mutation signatures of somatic point mutations were identified with the Bioconductor package SomaticSignatures (version 2.6.0) (30) using the non-negative matrix factorization (NMF) algorithm. Similarity between inferred signatures and the 30 previously identified signatures from Alexandrov *et al*. (31-33) was measured using *cosine* similarity.

MAF files were generated from all somatic mutations (excluding InDels) in coding regions and splice acceptor and donor sites for input to SomaticSignatures (v2.6.0) (30). Using SomaticSignatures, the frequencies of motifs were normalized to the frequency of 3-mers across the human genome. The non-negative matrix factorization (NMF) algorithm was used for decomposition, using 2-8 signatures and 20 replicates for each round. By visual inspection of plots of residual sum of squares (RSS) and explained variance vs the number of signatures, and similarity to the Alexandrov signatures, the optimal number of signatures was deemed to be six. For each signature, the *cosine* similarity to each for the 30 Alexandrov signatures (31, 32) was calculated, and the Alexandrov signature with the highest similarity was determined to be the best match (Supplementary Table S3). Signature C had highest *cosine* similarity to Alexandrov signatures 13 (.84) and 2 (.80) and is seen predominantly in case 21M. It is likely that both signatures are present in this sample, as they have previously been found together in other tumour types.

For comparison to sporadic CRC, somatic mutation calls from a total of 386 WES of colon adenocarcinomas (COAD) and rectum adenocarcinomas (READ) were downloaded from TCGA Data Portal (6 January 2017) (34). Four samples with less than 10 mutations were removed, leaving 382 samples. Protein coding and splice site mutations from the 382 samples, identified using the Variant Effect Predictor

software (12) as described above, were used to generate the CRC (TCGA) mutation spectrum in Fig. 3A. The Giannakis *et al*. (35) cohort included 619 samples from the Nurses' Health Study (NHS) and the Health Professionals Follow-up Study (HPFS). A complete list of somatic mutations from this cohort was downloaded and the protein coding and splice site mutations were used to generate the CRC (NHS/HPFS) mutation spectrum in Fig. 3A Similarity of the IBD-CRC mutation spectrum to the spectra derived from the Giannakis *et al*. (35) and TCGA cohorts were measured using *cosine* similarity.

Assessment of HPV Genome DNA in Case 21M

To search for HPV DNA, all sequencing reads from case 21M were remapped to a reference genome that consisted of the human reference genome GRCh37 plus 8321 viral genomes from NCBI Viral Genomes database (release 80; ftp://ftp.ncbi.nlm.nih.gov/refseq/release/viral/), which includes the high-risk human papillomavirus 16 and 18, plus 51 other HPV genomes. There were no read alignments to any HPV genomes, only a low level of poor alignments to repeat regions in other virus genomes. However, since WES sequencing was generated for this study, it is unlikely any viral sequence would be captured even if present.

HLA and Neo-epitope Analysis

HLA Class I 4-digit typing was performed using the HLA genotyping algorithm OptiType version 1.0 with default parameters (36). OptiType aligns the HLA sequence of the sample in question to HLA allele reference sequences and selects the optimal solution applying integer linear programming. To predict the number of neo-epitopes from missense mutations, the corresponding proteins were mapped to the GRCh37.74 human reference proteome. Peptides consisting of up to 17 amino

acids, depending on the position of the mutation in the protein (middle or termini), were retrieved and the affected wild-type amino acid was replaced *in silico* with the corresponding mutant amino acid for each mutant protein. Neo-epitope predictions were performed using the HLA-I Consensus algorithm of the Immune Epitope Database and Analysis Resource with 9mer sliding windows of the mutant peptides (37, 38). Prediction scores of all six HLA-I alleles and all sliding windows per mutant were considered in the analysis, therefore multiple neo-epitopes per mutation were possible to count. To cover most of the potential immune responses, neo-epitopes with a relative percentile rank ≤ 1% for each HLA-I allele were considered binders (39, 40).

In the present work the IEDB consensus tool for predicting HLA-peptide binding affinities was employed. The "consensus" approach involves the integration of several HLA-peptide binding prediction algorithm scores into a single score (http://tools.iedb.org/mhci/).

Since the different algorithms have different ways of "scoring" binding affinities, one way of integrating the different scores is by calculating percentile ranks.

A percentile rank represents a percentage of scores in a given distribution of scores, and it is calculated as follows:

$$Percentile\ rank = \frac{B + (0.5 * S)}{N} * 100$$

Where *B* is the number of scores below a given score, *S* is the number of occurrences of a given score, and *N* is the total number of scores in the distribution.

In this specific analysis, the HLA binding prediction tool generates a percentile rank for each peptide-HLA binding prediction method independently, by comparing a given

peptide-HLA binding score (e.g. IC50 value) to a set of predictions using random peptides from the SWISSPROT database (http://www.uniprot.org/uniprot/). Then, the median of the percentile ranks from the different methods for each peptide-HLA binding prediction is reported as the consensus percentile rank for that peptide-HLA prediction (http://tools.iedb.org/mhci/help/).

Finally, the consensus percentile rank of <1% was selected according to the IEDB recommended threshold for HLA-I predictions (http://help.iedb.org/hc/en-us/articles/114094151811-Selecting-thresholds-cut-offs-for-MHC-class-I-and-II-binding-predictions).

Nanostring Analysis

The NanoString nCounter Analysis System (NanoString Technologies, Seattle, WA, USA) was used to measure immune-related gene expression. Total RNA was diluted with RNase-free water to 20 ng/µl. We analyzed 100 ng (5 µl) of RNA from each sample using the PanCancer Immune Profiling Codeset (Nanostring Technologies). Each sample was analyzed in a separate multiplexed reaction. The CodeSet was hybridized in solution for 18 h at 65 °C according to the manufacturer's instructions. Hybridized samples were loaded onto the nCounter Prep Station for purification and immobilization followed by quantification of target RNA using the nCounter Digital Analyzer. Quantified expression data was imported into NanoString's nSolver Analysis Software (version 2.5) for quality checking and normalization. Raw counts were initially normalized using the internal positive controls permitting correction of potential sources of variation associated with the technical platform. Normalization for differences in RNA input was achieved using the 15 candidate housekeeping genes provided by Nanostring. Twenty-four of the cancer cases were analyzed to reflect

IBD-CRC heterogeneity (3Q, 4D, 5J, 6J, 7F, 8E, 9N, 10L, 12N, 14G, 15G1, 15G2, 16M, 17Z, 18F, 20Y, 21M, 22L, 23Y, 26N, 27N, 29D, 33W, 34S).

MLH1 Promoter Methylation Analysis

MLH1 promoter methylation was analyzed in a region spanning 17 CpG sites (41), from -169 to 465 of the human *MLH1* gene (NG_007109.2). CpG island methylation of this promoter region has been shown to correlate with MLH1 protein expression(41) and has been analyzed for *MLH1* promoter methylation assessment in colorectal cancer (42, 43). 76-200ng of tumour DNA was bisulfite-converted and purified using the EZ DNA Methylation Kit Gold (Zymo Research) according to manufacturer's protocol. HCT116 DNA served as a negative control and HCT116 DNA treated with CpG Methylase M.Sss1 (M0226S, New England Biolabs) was used as a positive control. Promoter methylation of the X-chromosomal *FMR1* gene, which is non-methylated in males and hemi-methylated in females(44), was assessed as an internal control for each sample. 1-4µl of bisulfite-treated DNA was amplified using GoTaq® G2 Hot Start Polymerase (M7405, Promega) and M13-tailed primers:

MLH1 forward

5' TGTAAAACGACGGCCAGTGGGAGGTTATAAGAGTAGGGTTAA '3'

and reverse

5'GCAGGAAACAGCTATGACCTCTCAACTCTATAAATTACTAAATCTCTT '3',

FMR1 [102] forward

5' TGTAAAACGACGGCCAGTTGAGTGTATTTTTGTAGAAATGGG '3'

and reverse

5' GCAGGAAACAGCTATGACCTCTCTCTTCAAATAACCTAAAAAC '3'. Cycling conditions (*MLH1/FMR1*) were 5min at 95°C, 5 cycles of 30sec at 95°C, 120sec at 55/60°C, 180sec at 72°, 37 cycles of 30sec at 95°, 60sec at 70/65°C, 180sec at 72°C, 15min at 72°C. Sanger sequencing using M13 forward primer

(5´TGTAAAACGACGGCCAGT'3´) was carried out with an Applied Biosystems 3130xl Genetic Analyzer (Life Technologies).

SYNCHRONOUS CANCERS IN IBD PATIENTS

Previous studies reported that synchronous colorectal cancers arise in IBD patients with a frequency of 11%-27% (45, 46) . Two of the patients in our series had multiple cancers: Patient 15G had two adjacent cancers (cases 15G1 and 15G2) and had a clinically significant germline variation in *CHEK2* and patient 32N had 3 cancers separated anatomically (cases 32N1, 32N2, 32N3) with an uncharacterised germline variation in *GALNT12* (Supplementary Table S5 and Supplementary Results).

Patient 15G was a male patient with an undefined period of colonic Crohn's disease who at the age of 65 years, developed two cancers separated by a distance of 10mm; case 15G1 showed an adenocarcinomatous pattern and 15G2 displayed a squamous cell carcinomatous differentiation. The two cancers from 15G had over 2410 common mutations suggesting that they originated from the same precursor clone. Both cancers have shared non-silent mutations in *MLH1* (I36S), *PMS2* (D544N), *POLE* (P286R and F348S in the exonuclease domain), *POLD1* (R81W) and *APOBEC3B* (A121T) but also private mutations in *POLE*, *POLD1*, *APC*, *MSH6*, *PMS2* and *APOBEC3B*. Neither have non-silent mutations in *KRAS* or *TP53* (Fig. 2). 15G2 has more private mutations (5646) than 15G1 (3170), however, in total 2410 mutations are shared between the two tumours suggesting that both evolved from the same precursor clone. Patient 15G also had a germline frameshift variation in *CHEK2* (T367fs, more commonly referred to in the literature as c.1100delC), which confers a risk of developing of breast, colon and prostate cancer (47) (Supplementary Table S5). It is unclear how chronic inflammation may modulate this

germline variant but as *CHEK2* is involved in the DNA damage response pathways functional deficiency may accelerate carcinogenesis (48).

Patient 32N is a female who was diagnosed with autoimmune hepatitis at a young age and had liver cirrhosis (proven on biopsy) the following year. She has been treated with continuous low dose steroid and azathioprine therapy to the present day. It is unclear when she developed ulcerative colitis (although she had intermittent diarrhoea for >15 years) and colorectal cancer was discovered at the age of 33 years. The colectomy specimen had 3 separate cancers from 3 different locations; case 32N1 (ascending colon) and case 32N2 (transverse colon) were adenocarcinomas and case 32N3 (descending colon) was a mucinous carcinoma. The cancers from 32N had only 6 somatic mutations common to at least 2 of the 3 cancers. No somatic mutations were common between all 3 cancers, demonstrating the inter-tumour heterogeneity and indicating that these 3 cancers have arisen independently and synchronously on a background of chronic inflammation (Fig. 2). Case 32N3 is the only cancer with a *MSH2* missense variant (Fig. 2), but the mutation signature in this tumour does not correlate with a dMMR mutational signature (Fig. 2), and it is not a hypermutator (Fig. 2). Germline analysis identified an uncharacterised *GALNT12* variation (G350R) in patient 32N. *GALNT12* is located on chromosome 9q22-33 in close proximity to a colorectal cancer linkage peak and *GALNT12* mutations have been described in patients that develop multiple epithelial cancers including mucinous colon and breast cancers (49, 50). *GALNT12* catalyses the transfer of an N-acetyl-D-galactosamine residue to a serine or threonine residue, one of the initial steps in O-linked oligosaccharide biosynthesis (glycosylation). Aberrant gylcosylation has been identified in several cancers including colorectal cancer (49, 51)

REFERENCES

1. Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., et al. A global reference for human genetic variation. Nature 2015, *526:* 68-74.
2. Li, H. and Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009, *25:* 1754-1760.
3. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010, *20:* 1297-1303.
4. Database of Single Nucleotide Polymorphisms (dbSNP). Bethesda (MD): National Center for Biotechnology Information, National Library of Medicine. (dbSNP Build ID: 135). http://www.ncbi.nlm.nih.gov/SNP/.
5. Tischler G, L. S. biobambam: tools for read pair collation based algorithms on BAM files. Source Code for Biology and Medicine. . pp. 2014;2019:2013. doi:2010.1186/1751-0473-2019-2013., 2014.
6. Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 2013, *31:* 213-219.
7. Database of Single Nucleotide Polymorphisms (dbSNP). Bethesda (MD): National Center for Biotechnology Information, National Library of Medicine. (dbSNP Build ID: 141). http://www.ncbi.nlm.nih.gov/SNP/.
8. Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res 2015, *43:* D805-811.
9. Saunders, C. T., Wong, W. S., Swamy, S., Becq, J., Murray, L. J., and Cheetham, R. K. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics 2012, *28:* 1811-1817.
10. Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016, *536:* 285-291.
11. Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. Integrative genomics viewer. Nat Biotechnol 2011, *29:* 24-26.
12. McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., et al. The Ensembl Variant Effect Predictor. Genome Biology 2016, *17:* 122.
13. Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., et al. Ensembl 2016. Nucleic Acids Res 2015, *44:* D710-716.
14. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. Initial sequencing and analysis of the human genome. Nature 2001, *409:* 860-921.
15. Maruvka, Y. E., Mouw, K. W., Karlic, R., Parasuraman, P., Kamburov, A., Polak, P., et al. Analysis of somatic microsatellite indels identifies driver events in human tumors. Nature Biotechnology 2017, *35:* 951.
16. Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res 2016, *44:* D862-868.
17. ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/.
18. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009, *25:* 2078-2079.
19. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 2011, *27:* 2987-2993.

20. OMIM Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD). World Wide Web URL: http://omim.org/.
21. Database of Single Nucleotide Polymorphisms (dbSNP). Bethesda (MD): National Center for Biotechnology Information, National Library of Medicine. (dbSNP Build ID: 147). http://www.ncbi.nlm.nih.gov/SNP/.
22. Favero, F., Joshi, T., Marquard, A. M., Birkbak, N. J., Krzystanek, M., Li, Q., et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. Ann Oncol 2015, *26:* 64-70.
23. Sheffer, M., Bacolod, M. D., Zuk, O., Giardina, S. F., Pincas, H., Barany, F., et al. Association of survival and disease progression with chromosomal instability: a genomic exploration of colorectal cancer. Proc Natl Acad Sci U S A 2009, *106:* 7131-7136.
24. The Cancer Genome Atlas, N. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012, *487:* 330-337.
25. Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013, *499:* 214-218.
26. http://www.broadinstitute.org/cancer/cga/mutsig_run.
27. Martincorena, I. i., Raine, K. M., Gerstung, M., Dawson, K. J., Haase, K., Van Loo, P., et al. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell 2017, *171:* 1029-1041.e1021.
28. Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., et al. A census of human cancer genes. Nature Reviews Cancer 2004, *4:* 177.
29. Mayer, C. Phobos. 3.3.11, 2006-2010.
30. Gehring, J. S., Fischer, B., Lawrence, M., and Huber, W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. Bioinformatics 2015, *31:* 3673-3675.
31. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., et al. Signatures of mutational processes in human cancer. Nature 2013, *500:* 415-421.
32. Alexandrov, L. B., Jones, P. H., Wedge, D. C., Sale, J. E., Campbell, P. J., Nik-Zainal, S., et al. Clock-like mutational processes in human somatic cells. Nat Genet 2015, *47:* 1402-1407.
33. http://cancer.sanger.ac.uk/cancergenome/assets/signatures_probabilities.txt.
34. The Cancer Genome Atlas homepage. http://cancergenome.nih.gov.
35. GiannakisÂ M., Mu, XinmengÂ J., Shukla, SachetÂ A., Qian, ZhiÂ R., Cohen, O., Nishihara, R., et al. Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma. Cell Reports 2016, *15:* 857-865.
36. Szolek, A., Schubert, B., Mohr, C., Sturm, M., Feldhahn, M., and Kohlbacher, O. OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics 2014, *30:* 3310-3316.
37. Kim, Y., Ponomarenko, J., Zhu, Z., Tamang, D., Wang, P., Greenbaum, J., et al. Immune epitope database analysis resource. Nucleic Acids Res 2012, *40:* W525-530.
38. Vita, R., Overton, J. A., Greenbaum, J. A., Ponomarenko, J., Clark, J. D., Cantrell, J. R., et al. The immune epitope database (IEDB) 3.0. Nucleic Acids Res 2015, *43:* D405-412.
39. Moutaftsi, M., Peters, B., Pasquetto, V., Tscharke, D. C., Sidney, J., Bui, H. H., et al. A consensus epitope prediction approach identifies the breadth of murine T(CD8+)-cell responses to vaccinia virus. Nat Biotechnol 2006, *24:* 817-819.

40. Kotturi, M. F., Peters, B., Buendia-Laysa, F., Jr., Sidney, J., Oseroff, C., Botten, J., et al. The CD8+ T-cell response to lymphocytic choriomeningitis virus involves the L antigen: uncovering new tricks for an old virus. J Virol 2007, *81:* 4928-4940.
41. Deng, G., Chen, A., Hong, J., Chae, H. S., and Kim, Y. S. Methylation of CpG in a small region of the hMLH1 promoter invariably correlates with the absence of gene expression. Cancer Res 1999, *59:* 2029-2033.
42. Gay, L. J., Arends, M. J., Mitrou, P. N., Bowman, R., Ibrahim, A. E., Happerfield, L., et al. MLH1 promoter methylation, diet, and lifestyle factors in mismatch repair deficient colorectal cancer patients from EPIC-Norfolk. Nutr Cancer 2011, *63:* 1000-1010.
43. Newton, K., Jorgensen, N. M., Wallace, A. J., Buchanan, D. D., Lalloo, F., McMahon, R. F., et al. Tumour MLH1 promoter region methylation testing is an effective prescreen for Lynch Syndrome (HNPCC). J Med Genet 2014, *51:* 789-796.
44. Boyd, V. L., Moody, K. I., Karger, A. E., Livak, K. J., Zon, G., and Burns, J. W. Methylation-dependent fragment separation: direct detection of DNA methylation by capillary electrophoresis of PCR products from bisulfite-converted genomic DNA. Anal Biochem 2006, *354:* 266-273.
45. Lam, A. K., Chan, S. S., and Leung, M. Synchronous colorectal cancer: clinical, pathological and molecular implications. World J Gastroenterol 2014, *20:* 6815-6820.
46. Greenstein, A. J., Slater, G., Heimann, T. M., Sachar, D. B., and Aufses, A. H., Jr. A comparison of multiple synchronous colorectal cancer in ulcerative colitis, familial polyposis coli, and de novo cancer. Ann Surg 1986, *203:* 123-128.
47. Huijts, P. E., Hollestelle, A., Balliu, B., Houwing-Duistermaat, J. J., Meijers, C. M., Blom, J. C., et al. CHEK2*1100delC homozygosity in the Netherlands--prevalence and risk of breast and lung cancer. Eur J Hum Genet 2014, *22:* 46-51.
48. Sohn, J. J., Schetter, A. J., Yfantis, H. G., Ridnour, L. A., Horikawa, I., Khan, M. A., et al. Macrophages, nitric oxide and microRNAs are associated with DNA damage response pathway and senescence in inflammatory bowel disease. PLoS One 2012, *7:* e44156.
49. Guda, K., Moinova, H., He, J., Jamison, O., Ravi, L., Natale, L., et al. Inactivating germ-line and somatic mutations in polypeptide N-acetylgalactosaminyltransferase 12 in human colon cancers. Proc Natl Acad Sci U S A 2009, *106:* 12921-12925.
50. Clarke, E., Green, R. C., Green, J. S., Mahoney, K., Parfrey, P. S., Younghusband, H. B., et al. Inherited deleterious variants in GALNT12 are associated with CRC susceptibility. Hum Mutat 2012, *33:* 1056-1058.
51. Tran, D. T. and Ten Hagen, K. G. Mucin-type O-glycosylation during development. J Biol Chem 2013, *288:* 6921-6929.

Figure Legends

Supplementary Figure S1: Exome sequencing fold coverage of inflammatory bowel disease associated colorectal cancer cases and matched normal lymph nodes.

Mean sequencing fold coverage (blue) of 34 Inflammatory Bowel Disease Colorectal Cancer (IBD-CRC) cases and matched normal lymph nodes (LN). WES yielded

between 45- and 90-fold coverage of the cancer samples and 26- to 93-fold coverage of the normal lymph node samples. Shown in red is the percentage of exome bases with less than 5-fold sequencing coverage.

Supplementary Figure S2 Mismatch repair protein expression analysis.

Cancer case 33W, a hypermutator, shows loss of expression of MLH1 (top left) and PMS2 (top right) and no loss of expression of MSH2 (bottom left) and MSH6 (bottom right) in the cancer tissue compared with the surrounding stromal tissue.

Supplementary Figure S3: CpG island methylation of *MLH1* promoter region.

Representative chromatograms of bisulfite-sequencing analysis of the *MLH1* promoter from -169 to -465, comprising 17 CpG islands. (**A**) Chromatogram of a non-methylated sample (case 16M) without any cytosines, indicating complete conversion of cytosine to uracil and thus no methylation. (**B**) Chromatogram of a methylated sample (case 27N) with asterisks indicating CpG islands, where cytosines either appear as Cs, or as C and T double peaks, indicating CpG island methylation.

Supplementary Figure S4: Rainfall plot for case 21M.

Somatic mutations are ordered along the horizontal axis according to genomic position. The intermutation distance is the genomic distance between a mutation and the previous neighbouring mutation. No clusters of mutations (as seen with kataegis) are evident in case 21M.

Supplementary Figure S5: Proportion of IBC-CRC genomes with SCNAs.

Shown are the total size of all genomic regions with predicted somatic copy number loss (blue), somatic copy number gain (red) in each tumour genome. Tumours are listed left to right in order of mutator status and decreasing mutation rate.

Supplementary Figure S6: Nanostring nCounter® pan-cancer immune panel analysis.

Euclidian distance based unsupervised hierarchical clustering of selected IBD-CRC cases, based on their expression values in the cytokine gene subset of the Nanostring nCounter® pan-cancer immune panel. The gene expression colour scale range represents under-expressed transcripts in blue and over-expressed transcripts in red. The top panel denotes the site of tumour, and the next panel describes mutator phenotype; hypermutator (red) and non-hypermutator (blue). Four hypermutated cases (6J, 4D, 26N and 15G1) cluster strongly.

Supplementary Figure S7: Validation of InDel mutations.

Shown are chromatograms generated from Sanger sequencing of regions with 1bp deletions in (a) *ACVR2A*, (b) *RPL22*, and (c) *RNF43* in tumour samples 28E, 27N and 33W, respectively, and their respective matched normal sample. The tumour samples have overlapping and shifted traces at the locations marked, which are not observed in the matched normal samples, indicating the presence of a somatic 1bp deletion in each tumour. The chromatograms have been cropped to show the relevant regions only.